贝叶斯混合效应模型：基于 brms 的应用教程

潘晚坷[1]　温秀娟[2]　金海洋[3]

([1]南京师范大学心理学院, 南京 210097; [2]广州医科大学附属脑科医院, 广州 510370; [3]Department of Psychology, Division of Science, New York University Abu Dhabi, PO Box 129188, Saadiyat Island, Abu Dhabi, United Arab Emirates)

通信作者: 金海洋, E-mail: haiyang.jin@outlook.com

摘　要：相比于传统方法，贝叶斯混合效应模型在处理数据层级结构和提供更直观的统计结果等方面有着独特的优势。这些诸多优势使其逐渐流行于心理学研究之中。然而，国内目前尚缺少将贝叶斯混合效应模型应用于心理学研究的介绍。因此，本文首先概述了贝叶斯混合效应模型的基本概念和原理，并且结合模拟数据展示了如何理解固定效应和随机效应，如何利用 R 语言的 *brms* 工具包设定和拟合贝叶斯混合效应模型，如何借助先验预测检验自定义的先验分布是否合理，以及如何使用贝叶斯因子进行假设检验。凭借其强大的灵活性，贝叶斯混合效应模型可以适用于多样的心理学研究。

关键词：贝叶斯，混合效应模型，分层模型，贝叶斯因子，brms

1　引言

近年来，心理学研究所使用的统计方法日渐丰富。随着新的统计方法的推广与普及，心理学研究者逐渐意识到了一些传统统计方法的局限性。作为实验心理学最常用的统计方法之一，重复测量方差分析（Repeated-measures analysis of variance）也不例外。首先，重复测量方差分析处理的是汇总后（aggregated）的数据，比如每个被试在每个条件下的平均反应时。而汇总后的数据忽略了每个被试个体内部的变异（variability），这可能会导致重复测量方差分析无法对自变量的效应进行准确的估计（Haaf & Rouder，2017；Veenman et al.，2022）。其次，重复测量方差分析通常只能考虑一个随机因素（random factor，Judd et al.，2017），而无法同时考虑多个随机因素。例如，重复测量方差分析一般把被试作为随机因素，借此通过被试样本的数据推断被试总体的效应。类似的，重复测量方差分析也可以把刺激材料作为随机因素，通过实验中使用的刺激材料推断刺激材料总体的效应。但重复测量方差分析无法同时考虑被试和刺激材料这两个随机因素。此外，当某个被试在某个条件下的汇总数据缺失时（例如在纵向研究中某些被试未能参加某个时间点的测试），重复测量方差分析通常只能删除存在缺失值的被试的所有数据，这会造成数据进一步的流失从而可能导致有偏差的估计（Magezi，2015；Matuschek et al.，2017）。

相比之下，线性混合效应模型（Linear Mixed-effects Models，LMM），亦称多水平模型（Multilevel Models）或层级模型（Hierarchical Models），可以更好地处理这些问题。混合效应模型可以考虑被试在各个条件内的变化性，数据中可能存在的相互依存关系和层级结构，并且能够更好地处理数据缺失等问题（Singmann & Kellen，2019；Tuerlinckx et al.，2006）。例如，在以学生为被试的实验中，学生来自于不同的班级，不同的班级属于不同的学校；同一个班级或同一个学校的学生的表现一般更相似。混合效应模型可以更好地考虑学生、班级和学校之间的层级关系，对实验效应进行更准确的估计。随着统计工具（比如 R 语言）的成熟，在心理学研究中使用混合效应模型来分析数据变得更加容易，越来越多的研究也开始使用该模型（Bono et al.，2021）。

与（频率论的，Frequentist）线性混合效应模型相比，贝叶斯线性混合效应模型（Bayesian Linear Mixed-effects Models，BLMM）因其在基于贝叶斯统计推断（即依照贝叶斯定理，通过结合先验信息（prior）与模型似然性（likelihood）对参数的后验（posterior）进行统计推断，详见下文 2.3）以及处理复杂模型等方面的优势受到了越来越多研究者的关注（Bürkner，2017；Sorensen & Vasishth，2016）。首先，与频率论统计相比，贝叶斯分析的一大优势是可以量化参数估计的不确定性（uncertainty），即贝叶斯分析的结果是一个分布（例如后验概率密度分布，详见 2.2），分布的宽窄描述了对参数分析的不确定性。其次，贝叶斯分析的可信区间（credible interval）可以被理解为一个拥有一定概率包含真值（true value）的区间。比如 95%的可信区间有 95%的可能性包含真值（Morey et al.，2016）。但该理解通常被误认为是频率论置信区间（confidence interval）的属性（Morey et al.，2016）（95%置信区间表示，在通过无数样本所获得的无数个区间中，大约有 95%的区间包含了真值（Lambert，2018））。其次，基于频率论的统计方法一般不能为支持虚无假设（null hypothesis）提供证据（Cohen，1994），而这些潜在的证据可以通过贝叶斯分析中的贝叶斯因子（Bayes Factor，BF）或实际等价区域（Region of Practical Equivalence， ROPE）获得（Makowski et al.，2019）。再者，当混合效应模型试图包含复杂的随机效应结构时，基于频率论的模型通常难以收敛（converge），进而无法给出可靠的估计（Matuschek et al.，2017）。相比之下，贝叶斯模型能够更好地处理复杂的随机效应结构（Sorensen & Vasishth，2016）。最后，贝叶斯分析还可以通过整合研究者的先验知识（例如研究者通过以往文献的结果获得的先验知识）提供更合理的参数估计。

贝叶斯混合效应模型兼具（频率论）混合效应模型和贝叶斯分析的优点，是心理学研究中极具潜力和值得推广的统计方法。本文致力于向读者提供一个相对容易理解和上手的教程。首先，我们将概述混合效应模型和贝叶斯统计的基本概念与原理。其次，借助模拟数据，展示如何使用统计编程语言 R 的 brms 工具包来进行贝叶斯混合效应模型的数据分析。最后，讨论贝叶斯混合效应模型在心理学研究中的拓展应用和一些常见问题。

2　模型介绍

1.1　混合效应模型

混合效应模型（Baayen et al.，2008）是一般线性模型的扩展形式，得名于它同时包含了固定效应（fixed effects）和随机效应（random effects）。固定效应通常是研究者所感兴趣的效应，即自变量在总体（population）水平的效应。与之相对的，随机效应是自变量在随机因素（例如被试或刺激材料）各个水平的特有效应。例如，在探究图片类型对情绪唤醒的影响的实验中，每个被试受到图片类型的影响可能并不完全相同。其中，图片类型对所有被试的平均影响为固定效应，而每个被试受到图片类型的影响与平均影响的差值为随机效应。

为了更好地解释固定效应与随机效应，我们首先回顾最基本的一般线性模型：

$$Y_{i,j} = \beta_0 + \beta_1 X_{i,j} + \epsilon_{i,j} \quad \text{公式 1}$$

其中，$Y_{i,j}$ 是因变量，比如被试的脑电波幅（详见 3 应用示例），表示第 j 名被试在第 i 试次的观测数据。$X_{i,j}$ 为自变量，表示研究者感兴趣的实验操作，可以是组间或者组内变量。参数 β_0 是截距，表示所有被试的均值，比如所有被试的平均脑电波幅；参数 β_1 为斜率，表示自变量对因变量的影响，比如不同图片类型对被试脑电波幅的影响（本文中涉及截距和斜率的解释仅适用于应用逐次差分对比编码（Successive Differences Contrast Coding，即本文所使用的编码）且只有两个水平的自变量（Venables & Ripley，2002）。当使用不同的编码时，截距和斜率的解释也会有所不同，详见 Schad 等，2020）。最后，$\epsilon_{i,j}$ 为残差（residual），表示模型中不能被解释的部分。

需要注意的是，使用一般线性模型需要满足残差独立性（independence）和正态性（normality）的前提预设（assumption），即不同观测数据的残差应该是相互独立并且服从正态分布。然而，在被试内实验设计中，每个被试完成不同图片类型的多个试次，因此他们在不同试次上的表现并不是完全独立的。此时使用一般线性模型无法处理观测数据间的依存性。此外，公式 1 的模型没有考虑不同被试间的个体差异，比如，不同被试的平均脑电波幅（即每个被试的截距）可能存在差异，而忽略这些差异可能会掩盖图片类型的真实效应。为了更好地考虑这些方面，可以在公式 1 的基础上增加随机截距：

$$\begin{aligned} Y_{i,j} &= \beta_{0,j} + \beta_1 X_{i,j} + \epsilon_{i,j} \\ \beta_{0,j} &= \beta_0 + S_{0,j} \end{aligned} \quad \text{公式 2}$$

与公式 1 相比，公式 2 主要区分了固定截距（β_0）和随机截距（$S_{0,j}$）。其中，固定截距可以被理解为所有被试截距的平均值，即所有被试的平均脑电波幅；而随机截距可以被理解为每个被试的截距与固定截距的差值，即每个被试各个条件的平均脑电波幅与所有被试平均脑电波幅的差值。虽然加入随机截距可以考虑到各被试平均脑电波幅的差异，但仍然没有考虑到自变量对因变量的影响在不同被

试上的差异。比如，不同被试受到图片类型的影响并不一定相同。为了更好地解释自变量的影响在不同被试间的差异性，可以进一步在混合模型中加入随机斜率：

$$\begin{aligned} Y_{i,j} &= \beta_{0,j} + \beta_{1,j} X_{i,j} + \epsilon_{i,j} \\ \beta_{0,j} &= \beta_0 + S_{0,j} \\ \beta_{1,j} &= \beta_1 + S_{1,j} \end{aligned} \quad \text{公式 3}$$

与公式 2 相比，公式 3 主要区分了固定斜率（β_1）和随机斜率（$S_{1,j}$）。其中，固定斜率是所有被试斜率的平均值，即所有被试受到图片类型的平均影响；随机斜率是每个被试斜率与平均斜率的差值，即每个被试受到图片类型的特有影响。需要注意的是，由于随机截距与随机斜率都受到同一个随机因素（被试）的影响，因此两者可能存在一定的相关性，比如对于平均脑电波幅（随机截距）更高的被试，其在不同图片类型下的脑电波幅的差异（随机斜率）可能更大。此时通过预设两种随机效应来自同一个多元正态分布可以更好地考虑他们之间的相关性。

总的来说，混合效应模型是一般线性模型的扩展，它能够借助随机截距和随机斜率考虑随机因素（例如被试和实验材料）对实验效应的影响，可以更好地处理数据中的依存和层级关系，弥补重复测量方差分析的不足，以提高对于实验效应估计的准确性。

1.2 贝叶斯混合效应模型

在构建混合效应模型过程中，为了在控制合理的一类错误率的同时保持良好的统计效力，研究者通常需要考虑最大化的模型（maximal model，Barr et al.，2013），即把所有可能的随机截距、随机斜率和它们之间的相关系数加入模型中。然而，由于模型参数增多，这可能导致模型难以收敛，从而给出不可靠的估计（Bates et al.，2015）。相比之下，贝叶斯混合效应模型可以更好地处理这种复杂的模型。更重要的是，相较于频率论混合效应模型，贝叶斯混合效应模型具有考虑先验信息和提供更直观的参数后验分布等其他更多优势。这些优点使得贝叶斯混合效应模型更受欢迎。

贝叶斯定理（Bayes’ theorem）在贝叶斯数据分析中起着重要的作用，它借助数据更新先验信息，从而得到后验信息：

$$p(\theta|Y) = \frac{p(Y|\theta) * p(\theta)}{p(Y)} \quad \text{公式 4}$$

其中，$p(\theta|Y)$ 为参数的后验分布（posterior distribution，即连续型参数对应的概率密度函数或离散型参数对应的概率质量函数），表示在观测到数据 Y 的基础上参数 θ 的概率分布。对于贝叶斯混合效应模型来说，θ包括所有的固定效应β、随机效应S、随机效应间的相关性和残差ϵ（见公式 3）所对应的分布的所有参数。$p(\theta|Y)$ 可以由公式右侧部分计算得来，其中 $p(\theta)$ 为参数的先验分布（prior distribution），表示研究者在观测到数据前对参数 θ 的先验知识。$p(Y|\theta)$ 为似然（likelihood），描述了当参数 θ 为不同的值时，观测到当前数据的似然值。值得注意的是，$p(Y)$ 并不包含 θ，这是因为 $p(Y)$是一个边际似然（marginal likelihood），它是通过对所有可能的θ的似然值与相应的先验概率乘积

求和或积分而获得（当 θ 为离散变量时：$p(Y)=\sum_\theta p(Y|\theta)*p(\theta)$；当 θ 为连续变量时，$p(Y)=\int_\theta p(Y|\theta)*p(\theta)d\theta$，（Kruschke & Liddell，2018）。$p(Y)$是一个“归一化”因子（normalizing factor），能够确保后验分布是一个有效的概率分布（使后验分布的面积为 1）。

基于贝叶斯定理，在对贝叶斯混合效应模型进行参数估计时，研究者首先需要为各参数设置合理的先验分布，之后通过数据对先验进行更新，从而获得参数的后验分布。需要注意的是，随着模型变得复杂，贝叶斯定理中的边际似然（公式4右侧分母部分）将变得难以计算。此时通常通过马尔科夫链蒙特卡洛（Markov Chain Monte Carlo，MCMC，Lambert，2018；Morey et al.，2011）算法推测参数的后验分布（Bürkner，2017），其具体原理可参考 Plummer（2023）。与 MCMC 相关的重要参数会在下文中结合示例予以介绍（3.2 节）。

3 应用示例

我们将借助一个假想的心理学实验展示如何使用 brms 工具包进行贝叶斯混合效应模型的数据分析。具体来说，本节将从数据模拟和预处理、模型的构建、先验的设定、模型的拟合与检验和统计推断几个方面介绍贝叶斯分析最基本的工作流程（Bayesian workflow，Gelman et al.，2020）。具体代码请见 https://osf.io/cgxwn/。

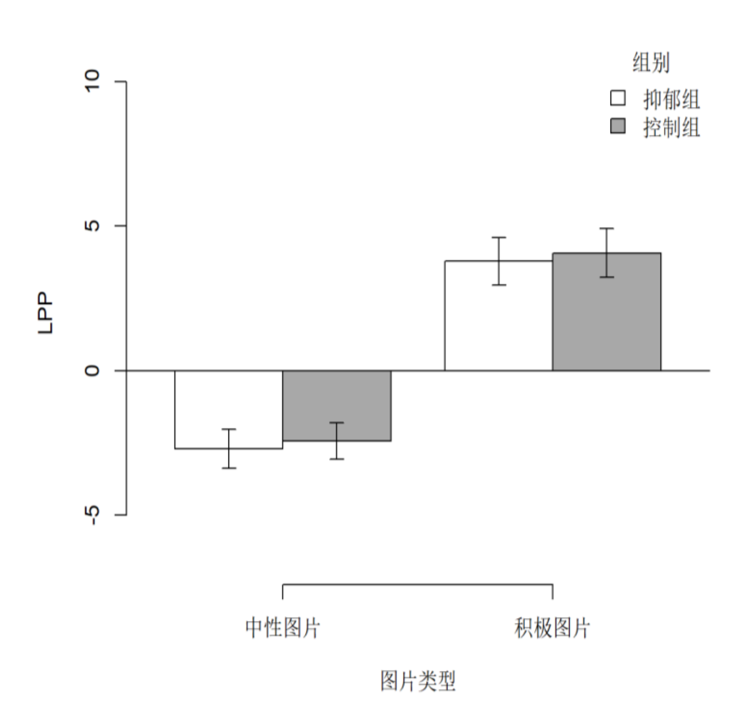


图 1 模拟数据柱状图

1.3 数据模拟与预处理

该假想研究的目的是探究抑郁症人群加工不同情绪刺激的神经基础。在该实验中，40 名抑郁症患者和 40 名健康对照组被试观看 30 张积极和 30 张中性图片，期间我们采集了他们的脑电数据。我们所感兴趣的因变量是晚期正电位（late positive potentials，LPP）的波幅。该实验为 2（组别 group：抑郁症组 depression、对照组 control）× 2（图片类型 type：积极 positive、中性 neutral）混合实验设计，其中组别为被试间因素，图片类型为被试内因素。

该假想实验的数据使用 R 的 *faux* 工具包生成（DeBruine，2021），具体代码请见补充材料数据模拟部分（https://osf.io/3g79b）。简单来说，我们设定（图 1 和表 1）图片类型的主效应为 6.5（即积极图片诱发的 LPP 比中性图片高 6.5），组别的主效应为 0.1（即控制组 LPP 比抑郁组高 0.1）和组别与图片类型的交互作用为 0.1。

表 1 模拟数据表述统计结果

组别	图片类型	总体参数(均值)	均值	标准差
抑郁组(n=40)	中性图片	-2.70	–2.71	2.12
	积极图片	3.83	3.78	2.56
控制组(n=40)	中性图片	-2.55	–2.43	1.97
	积极图片	4.12	4.07	2.62

注：“总体参数”为生成模拟数据时设定的参数；“均值”和“标准差”为模拟数据的描述性统计结果。

首先，我们需要确保数据集的结构符合工具包的要求。重复测量方差分析通常需要“宽”（wide）格式的数据，即每一行是一个被试的数据。相比之下，混合效应模型的分析通常需要 “长”（long）格式的数据，即每一行是一个试次的数据。例如，R 语言中的 *brms* 工具包需要使用长数据格式进行混合效应分析。与大部分实验程序（例如 E-prime）生成的数据类似，我们通过 *faux* 工具包生成的模拟数据也是以“长”格式保存，符合 *brms* 工具包输入数据的要求。该模拟数据集的变量名为 *df_simu*，我们可以通过 *head(df_simu, 10)*查看前 10 行数据：

```
head(df_simu, 10) # 查看数据结构
## A tibble: 10 x 4
##    subj         group      type         LPP
##    1 s01     depression   netural            -1.64
##    2 s01     depression   netural            -4.00
##    3 s01     depression   netural            -1.53
##    4 s01     depression   netural            -3.22
##    5 s01     depression   netural            -4.34
##    6 s01     depression   netural             1.49
```

```
## 7 s01 depression netural -3.46
## 8 s01 depression netural 2.93
## 9 s01 depression netural -3.27
## 10 s01 depression netural -1.44
```

此外，我们还需要对数据集 *df_simu* 中的称名（自）变量（nominal factors）进行编码（contrast coding）。这里我们使用逐次差分对比编码（Successive Differences Contrast Coding）对组别（group）和图片类型（type）这两个自变量进行编码，即使用 R 中的 *MASS::contr.sdif()* 函数。通过此设定，贝叶斯模型的输出结果刚好对应大部分研究者熟悉的主效应和交互作用（见公式 1 处解释）。

```
# 设定-0.5 和 0.5 的编码
contrasts(df_simu$group) <- MASS::contr.sdif(2)
contrasts(df_simu$type) <- MASS::contr.sdif(2)
```

1.4 构建模型

完成数据的预处理后，研究者需要构建模型。模型构建主要有两种方法。一种是从包含了研究者最感兴趣效应的简单模型开始，逐渐添加更多的效应以寻找能够最优地解释数据的模型。另外一种方法是根据实验设计构建最大化的模型（Barr et al.，2013）。鉴于第二种方法相对简单，且更多的被认知和心理学研究者所采用（Schad et al.，2021），本文将主要介绍如何根据实验设计构建最大化模型（对第一种途径感兴趣的读者可参考（Schad et al.，2021）。

在实际分析中建立贝叶斯线性混合效应模型时，研究者可以直接根据实验设计设定最大化的模型。需要注意的是，本文为了更清晰地介绍如何理解和设定固定效应和随机效应，我们将以三个模型为例，分别解释如何结合实验设计使用 *brms* 设定固定效应、随机截距和随机斜率。

我们首先建立一个仅包含固定效应的模型（*model_1*）。具体来说，我们将组别（*group*）和图片类型（*type*）两个自变量以及他们的交互作用作为固定效应。该模型假定所有被试的平均波幅都是相同的，且图片类型对于每个被试的影响也是相同的。

```
# 建立只有固定效应的模型
model_1 <- brm(
  LPP ~ group * type,
  data = df_simu,
  chains = 4,
  warmup = 1000,
  iter = 5000,
  thin = 1,
  cores = 4
)
```

在上述代码中，*LPP* 为因变量，即被试观看不同图片时 LPP 的波幅，*group * type* 表示组别 group 和图片类型 type 两个自变量的主效应以及他们的交互作用。*df_simu* 为数据集。其他参数是模型拟合时所基于的 MCMC 采样算法的相关参数。为了保证后验分布的有效性，MCMC 通常会同时运行多个独立的采样过程，每个独立的采样过程称为一条链（chain）。只有当各个独立链的采样“融合”在一起时（详见下文），MCMC 获得的参数后验分布才更可能是有效的（McElreath，2020）。在 *model_1* 中，我们设置了 4 条链（*chains = 4*，默认为 4）。与之相对应的是 *cores* 参数，它可以设定运行 MCMC 时使用的计算机中央处理器核心（Central Processing Unit cores）的个数。设置 *cores=4*（默认为 1）表示使用 4 个中央处理器核心同时分别各运行 1 条马尔可夫链（共 4 条链），借此提高数据分析的速度。*iter* 参数（iteration 的简写）指的是每条链的迭代次数，即对每个参数进行多少次的采样（sample）。*warmup* 参数（或称为 *burn in*）设定了在每条链采样的开始阶段用于“校准”采样器（sampler）以提升采样效率的迭代次数（详见 https://mc-stan.org/docs/reference-manual/hmc-algorithm-parameters.html），该阶段的采样不会被纳入最终的参数概率分布。*thin* 参数为“稀释比例（thinning rate，默认为 1）”，表示每隔 *thin* 个采样保留一个样本，研究者可以通过设定较大的 thin 来减少电脑内存的需求（详见 https://mc-stan.org/docs/reference-manual/effective-sample-size.html#thinning-samples）。总的来说，每条链最后保留的采样数量与 *iter*、*warmup* 和 *thin* 相关，最后保留的采样数等于（*iter* -*warmup*）/*thin*。例如，在 *model_1* 中每一条链会保留每个参数的 4000 个样本，即（5000-1000）/1，最终 4 条链一共会保留 16000 个样本。

通过 *summary()* 函数可以展示模型拟合的相关参数和结果：

```
summary(model_1) # 可以查看参数情况
##  Family: gaussian
##   Links: mu = identity; sigma = identity
## Formula: LPP ~ group * type
##    Data: df_simu (Number of observations: 4800)
##   Draws: 4 chains, each with iter = 5000; warmup = 1000; thin = 1;
##          total post-warmup draws = 16000
##
## Population-Level Effects:
##                  Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept            0.68      0.04     0.59     0.76 1.00    19010    12439
## group2M1             0.28      0.09     0.11     0.45 1.00    21511    12581
## type2M1              6.50      0.09     6.33     6.67 1.00    22435    12931
## group2M1:type2M1     0.02      0.17    -0.32     0.35 1.00    20521    11983
##
## Family Specific Parameters:
```

```
##        Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sigma      3.03      0.03     2.97     3.09 1.00    20821    12632
##
## Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
## and Tail_ESS are effective sample size measures, and Rhat is the potential
## scale reduction factor on split chains (at convergence, Rhat = 1).
```

输出结果的第一部分展示了模型设置的信息。例如，“Family：gaussian，Links: mu = identity; sigma = identity”表示模型预设自变量和因变量间的关系是线性的，且残差是一个正态分布。对于其他非线性关系的模型，研究者可以设定相应的残差分布和链接函数（详见第 4 部分）。

Population-Level Effects 部分为固定效应的结果，通常是研究者最为关注的部分。其中，*Intercept* 是模型的固定截距，表示所有被试的平均 LPP 波幅。*group2M1* 表示在平均图片类型的效应之后，健康对照组的 LPP 波幅减去抑郁症组 LPP 波幅的差值，即组别的主效应。与此类似，*type2M1* 是在平均组别的效应之后，积极图片诱发的 LPP 波幅减去中性图片诱发的 LPP 波幅的差值，即图片类型的主效应。*group2M1:type2M1* 是组别和图片类型交互作用，即 $\left(LPP_{\text{健康对照组,积极图片}} - LPP_{\text{健康对照组,中性图片}}\right) - \left(LPP_{\text{抑郁症组,积极图片}} - LPP_{\text{抑郁症组,中性图片}}\right)$。

在输出的结果中，*Estimate* 表示参数的后验分布的均值，*Est.Error* 表示后验分布的标准差（standard deviation）。*l-95% CI* 和 *u-95% CI* 表示参数的 95%可信区间的两个边界，例如 *group2M1* 的 95%的可信区间是[0.11，0.46]。*Rhat* 表示不同链条采样的融合程度。*Rhat* 越接近于 1，表示不同链条采样融合得更好，即模型的收敛情况更好。如果 *Rhat* 大于 1.1（或更严格的标准下，如大于 1.01）则说明模型的收敛情况并不理想（Gelman & Rubin，1992；Vehtari et al.，2021）。此时，研究者不应该使用获得的后验分布进行进一步的分析，而应该检查模型设定和数据是否存在问题，并进行相应的调整。*ESS* 为有效样本量（effective sample size），是对 MCMC 算法采样到的有效且独立的样本数量的估计。*ESS* 越高表明能用来有效估计后验分布的样本数量越大。通常建议 *Bulk-ESS* 应是 MCMC 链数量的 100 倍及其以上。例如，当使用 4 条 MCMC 链时，*Bulk-ESS* 应至少为 400（Vehtari et al.，2021）。为了以确保足够的 *Bulk-ESS*，研究者可以适当增加 *iter*，例如本文使用的 *iter* 为 5000。*Tail-ESS* 表示后验分布两端的有效样本量。通常情况下，过小的 *Bulk-ESS* 和 *Tail-ESS* 会伴随着过大的 *Rhat*。设定合理的模型和先验分布一般可以解决大部分 *Rhat* 过大的情况。最后，研究者还可以通过可视化的方式检查各条链的融合情况：

```
mcmc_trace(model_1, pars = c("b_Intercept", "b_group2M1", "b_type2M1", "b_group2M1:type2M1", "sigma"))
```

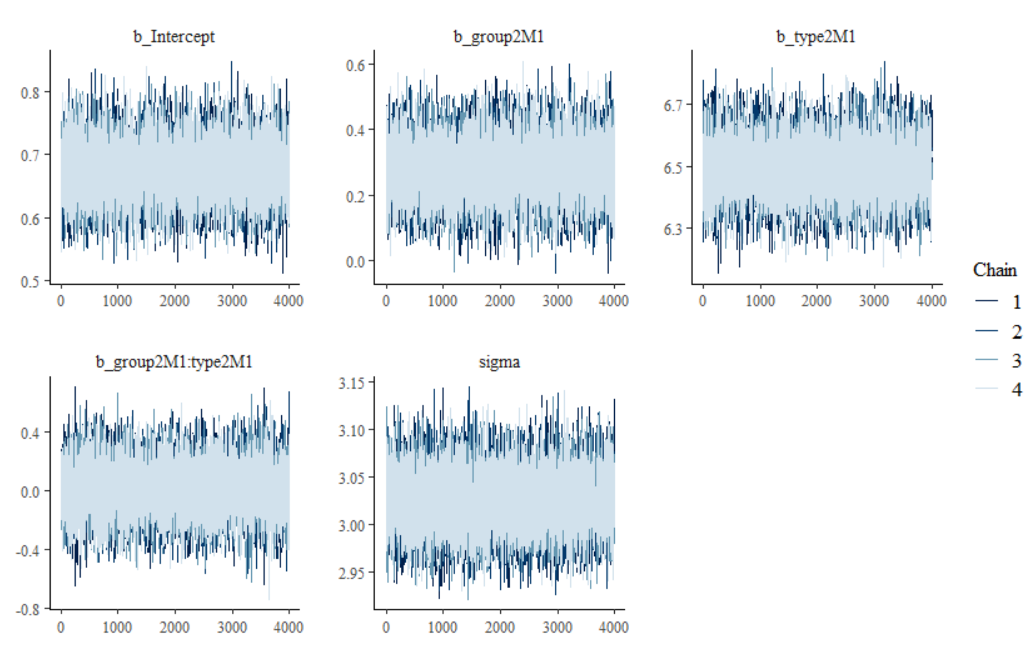


图 1 model_1 的 MCMC 各条链采样轨迹图（又称“毛毛虫”图）

注：x 轴表示迭代数，y 轴表示参数值。不同颜色的线表示不同的链。

图 2 描述了 *model_1* 中 MCMC 不同链抽样的轨迹图，反映了不同链的融合情况。在图 1 中，不同链参数的抽样很好地 “融合”在一起，说明其对应的参数后验分布是相对可靠的。由于融合的链看起来像一条毛毛虫，因此图 1 的轨迹图又称“毛毛虫”图。如果一条链的抽样与其他链的抽样存在一定的差异（通常也会伴随着较大的 *Rhat*），那这些后验抽样也很可能无法有效地表征参数的后验分布（详见补充材料第三部分 MCMC 链不收敛示例，https://osf.io/3g79b），此时，研究者需要重新考虑模型的合理性（Roy，2020）。

如前所述，仅包含固定效应的模型（*model_1*）假定所有被试的平均脑电波幅都是相同的，即所有被试共享同一个截距值。但这种情况在现实中不太可能发生。更可能的情况是，每个被试的平均脑电波幅都各不相同。因此，我们需要在模型中增加随机截距来考虑被试在平均脑电波幅上的个体差异。具体来说，我们需要在 *model_1* 的基础上增加基于被试（by-subject）的随机截距（对应公式 2 中的 $S_{0,i}$），即 *(1 | subj)*。其中 *subj* 为被试编号，*1* 表示随机截距：

```
model_2 <- brm(
  LPP ~ group * type + (1 | subj),
  data = df_simu,
  chains = 4,
  warmup = 1000,
  iter = 5000,
```

```
  thin = 1,
  cores = 4
)
summary(model_2) # 查看参数情况 （部分输出结果）
## Group-Level Effects:
## ~subj (Number of levels: 80)
##               Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sd(Intercept)     2.15      0.17     1.84     2.52 1.01      934     1946
##
## Population-Level Effects:
##                     Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept               0.67      0.24     0.19     1.15 1.00      502     1018
## group2M1                0.26      0.49    -0.72     1.20 1.01      450     1095
## type2M1                 6.50      0.06     6.38     6.62 1.00    13087    11075
## group2M1:type2M1        0.02      0.13    -0.23     0.26 1.00    13188    10601
```

Population-Level Effects 为 *model_2* 固定效应的结果，其相关参数的含义与 *model_1* 中的相同（详见 3.2 部分）。*Group-Level Effects* 为随机效应结构的结果，其中 *sd(Intercept)* 为随机截距的标准差，反映了 LPP 平均波幅在不同被试间的离散程度。虽然 *model_2* 考虑了不同被试平均 LPP 波幅的个体差异，但它仍然假定所有被试受到实验处理（即图片类型）的影响是相同的。然而在实际情况中，每个被试受到图片类型的影响更可能是不同的。因此，我们需要设置随机斜率考虑图片类型对不同被试的不同影响。具体来说，我们可以在 *model_2* 模型的基础之上加入随机斜率（对应公式 3 中的$S_{1,i}$），即 *(type | subj)*，其中 *subj* 为被试编号，*type* 为图片类型的随机斜率。与此同时，我们也在模型中加入了随机截距与随机斜率间的相关:

```
model_3 <- brm(
  LPP ~ group * type + (1 + type | subj),
  data = df_simu,
  chains = 4,
  warmup = 1000,
  iter = 5000,
  thin = 1,
  cores = 4
)
summary(model_3) # 查看参数情况 （部分输出结果）
# Group-Level Effects:
#   ~subj (Number of levels: 80)
#                          Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
# sd(Intercept)                2.17      0.18     1.86     2.55 1.00     1151     2404
```

```
# sd(type2M1)                   1.82      0.16      1.53      2.16 1.00      2783      5543
# cor(Intercept,type2M1)        0.32      0.11      0.10      0.52 1.00      1737      3516
#
# Population-Level Effects:
#                          Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
# Intercept                     0.68      0.24      0.19      1.14 1.01       623      1239
# group2M1                      0.27      0.47     -0.64      1.21 1.01       630      1343
# type2M1                       6.49      0.21      6.08      6.90 1.00      1535      3228
# group2M1:type2M1              0.00      0.42     -0.84      0.84 1.00      1595      3085
```

与 *model_2* 结果相比，*model_3* 中的 *Group-Level Effects* 部分额外展示了基于被试的随机斜率和它与随机截距相关性的后验分布情况。其中，*sd(type2M1)* 为基于被试的图片类型随机斜率的标准差，反映了不同被试 LPP 波幅受到图片类型不同影响的离散程度。*cor(Intercept, type2M1)* 为随机截距和随机斜率的相关，该相关结果表明两者之间存在一定的正相关，即平均 LPP 波幅更大的被试，其对积极图片的 LPP 波幅也比中性图片更大。

至此，我们完成了对该模拟数据的最大化模型的构建，即根据实验设计设定了所有合理的固定效应和随机效应结构。在该示例中，我们从仅包含固定效应的线性模型出发，在不同的模型中依次展示了如何添加随机截距和随机斜率。需要强调的是，在实际的贝叶斯数据分析中，研究者可以根据实验设计直接设定最大化的模型，而无需在不同的模型中依次添加不同的随机效应。

如前所述，贝叶斯混合效应模型还可以同时考虑多个随机因素，以及不同随机因素间的层级关系。例如，在分析包含“学生-班级-学校”层级关系的数据时，研究者可以通过设定“（1丨学校\班级\学生）”的随机效应结构来考虑同一个班级和同一个学校学生间的相似性等（详见，Bürkner，2018）。

1.5 先验的设置

完成构建模型后，研究者需要根据当前的先验知识自定义模型的先验，从而在数据分析中整合研究者的先验知识，这也是贝叶斯分析的一大优势。由于我们在先前建立的模型（例如 *model_3*）中没有自定义先验分布，因此这些模型使用的均是 *brms* 预先设定的默认先验分布。这些先验分布通常是弱信息先验（weakly informative prior），即相对实验数据来讲，基本不能为参数估计提供有效信息的先验（Seaman et al.，2012）。然而，使用这些默认的弱信息先验不仅不能发挥贝叶斯数据分析的最大作用，甚至可能会对数据分析的结果产生负面的影响（Seaman et al.，2012）。例如，贝叶斯因子很容易受到先验分布的影响，不合理的先验会导致有偏差的贝叶斯因子。因此，为贝叶斯分析设定合理的先验至关重要。

作为设定先验分布的第一步，研究者需要了解自己可以对一个模型的哪些参数设置先验。这里可以借助 *get_prior()* 函数得到相应的信息：

```
get_prior(
  LPP ~ group * type + (1 + type | subj),
  data = df_simu
)
##                 prior     class              coef group resp dpar nlpar lb ub
##                (flat)         b
##                (flat)         b          group2M1
##                (flat)         b group2M1:type2M1
##                (flat)         b           type2M1
##                lkj(1)       cor
##                lkj(1)       cor                    subj
##  student_t(3, 0.5, 5) Intercept
##    student_t(3, 0, 5)        sd                                        0
##    student_t(3, 0, 5)        sd                    subj                0
##    student_t(3, 0, 5)        sd         Intercept  subj                0
##    student_t(3, 0, 5)        sd           type2M1  subj                0
##    student_t(3, 0, 5)     sigma                                        0
```

如结果所示，我们可以为 *model_3* 模型的以下参数设定先验：固定截距（*class* 为 *Intercept*）；固定斜率（*class* 为 *b*），包含组别主效应、图片类型主效应和他们的交互作用（分别对应 *coef* 为 *group2M1*, *type2M1* 和 *group2M1:type2M1*）；基于被试（*group* 为 *subj*）的随机效应的标准差（*class* 为 *sd*），包含随机截距（*coef* 为 *Intercept* ）和随机斜率（*coef* 为 *type2M1*），基于被试的随机效应间的相关（*class* 为 *cor*），以及残差（*class* 为 *sigma*）。由于 *sd* 和 *sigma* 不能为负数，因此这些参数的最小值为 0（即列 *lb*，为 low bound 的缩写）。

在实际数据分析中，研究者一般可以根据以往研究的结果为当前研究的模型设定先验分布。此时，研究者可以通过先验预测检验（prior predictive check）来测试这些先验的合理性，并进行相应的调整，进而为模型确定合理的先验。在进行先验预测检验时，需要考虑领域知识（domain knowledge），即研究者对研究领域的专业知识和经验，它可以帮助研究者确定先验分布的合理范围，并检查先验分布是否与领域知识一致。如果先验分布与领域知识相矛盾，则需要重新考虑先验分布的选择。

如果可以从先前类似的研究获得参数的后验概率分布，研究者可以将这些概率分布作为当前研究模型的先验分布。即使无法获得参数准确的概率分布，研究者仍然可以根据先前研究结果以及自己已有的知识为模型设定先验分布。假设我们无法从先前研究中获取参数的概率分布作为模型的先验。此时，我们可以首先尝试设置弱信息的先验分布：

```
## 定义先验分布
prior_01 <- c(
```

```
  prior(normal(0, 10), class = Intercept),
  prior(normal(0, 10), class = b),
  prior(normal(0, 5), class = sigma),
  prior(normal(0, 5), class = sd),
  prior(lkj(2), class = cor)
)
```

在上述代码中，我们把固定截距和固定斜率分别设定为一个均值为 0，标准差为 10 的正态分布（在代码中表示为 *normal(0, 10)*）。对于 *prior(normal(0, 10), class = Intercept)*，我们可以简单地把它理解为，我们设定所有被试的平均值脑电波幅大致有 95%的概率落在-20 和 20 之间（即两个标准差之内）。对于平均脑电波幅来讲，这是一个相对比较大的区间。因此，它对应的先验分布是弱信息先验分布。类似地，所有固定斜率，包含组别和图片类型的主效应以及他们的交互作用，也大致落在相同的区间。此外需要注意的是，尽管我们使用正态分布作为残差 *sigma* 和随机效应的标准差 *sd* 设置先验，但由于这些参数不能小于 0，因此他们实际对应的先验分布是截断正态分布（truncated normal distribution）。最后，随机效应间的相关性设置为弱信息先验 *prior(lkj(2), class = cor)*，其中 *lkj* 为 LKJ 相关性矩阵（LKJ-Correlation prior matrix，Lewandowski et al.，2009）。

为了检验先验分布的合理性，我们可以使用先验预测检验考查仅基于先验信息生成的因变量（即 LPP 的波幅）的情况。进行先验预测检验需要对先验分布进行采样（sample），即在模型中设定先验分布 *prior=prior_01* 的基础上加入 *sample_prior ="only"*，通过该设置可以仅从先验分布中进行采样（忽略观测数据的似然值信息）。其余参数的设定和我们在分析数据时设定相同（例如 *model_3*），具体代码如下：

```
model_ppc_01 <- brm(
  LPP ~ group * type + (1 + type | subj),
  data = df_simu,
  chains = 4,
  warmup = 1000,
  iter = 5000,
  thin = 1,
  cores = 4,
  prior = prior_01,
  sample_prior = "only"
)
```

通过上述代码，我们可以获得 16,000 组（（5000-1000）*4）来自先验分布的抽样。其中，每一组都包含模型的所有参数，即固定截距，固定斜率和随机效应等。每一组先验参数都可以生成一套新的与数据集（*df_simu*）结构相同的模拟观测数据（即因变量）。我们可以通过考查这些模拟观测数据的合理性（比如检查每一套模拟数据的极值）去判断先验的设定是否合适。例如在一个以反应时为因变量的 stroop 实验中，如果那些由先验分布生成的模拟数据集中大部分数据集的最大值都超过预期（比如

10 秒），这很可能说明模型某些参数先验的设定不太合理（比如截距或标准差过大）。在当前模拟实验中，尽管我们无法根据先前研究获取精确的先验，但是可以利用现有的知识，即试次水平的脑电事件相关电位的波幅通常在一定的范围之内，来确定先验。因此，我们计算 16,000 套由先验分布生成的模拟观测数据各自的最小值（min）、均值（mean）和最大值（max），并通过它们的范围是否符合我们的先验知识来判断先验分布的合理性。

上述分析可以通过 *brms* 提供的 *pp_check()* 函数实现。以下代码展示了 *prior=prior_01* 的先验预测的结果：

```
pp_check(model_ppc_01, type = "stat", stat = "min", prefix = "ppd")
pp_check(model_ppc_01, type = "stat", stat = "max", prefix = "ppd")
pp_check(model_ppc_01, type = "stat", stat = "mean", prefix = "ppd")
```

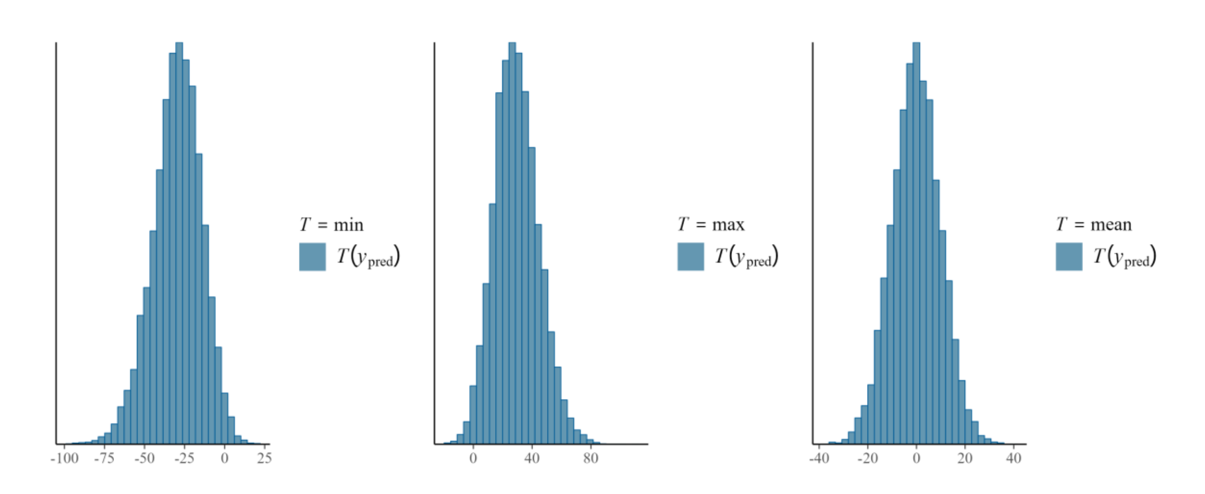


图 3 基于先验 prior_01 生成的 16,000 套预测数据（y(pred)）的最小值、最大值和均值的分布

结果显示，基于 *prior_01* 生成的每一套模拟观测数据的最小值大部分落在[-75, 0]，均值大部分落在[-25, 25]，而最大值大部分落在[0, 80]。根据先验知识，试次水平的脑电波幅通常很少会低于-50（μV）或高于 50（μV）（伪迹除外）。因此，我们可以适当地调整参数的先验分布以符合我们的先验知识，即减小参数先验分布的不确定性。

在接下来的尝试中，我们将固定截距和固定斜率的先验设置成均值为 0，标准差为 5 的正态分布，并进行与之前相同的先验预测检验：

```
## 定义先验分布
prior_02 <- c(
  prior(normal(0, 5), class = Intercept),
  prior(normal(0, 5), class = b),
  prior(normal(0, 5), class = sigma),
  prior(normal(0, 5), class = sd),
  prior(lkj(2), class = cor)
```

```
)
model_ppc_02 <- brm(
  LPP ~ group * type + (1 + type | subj),
  data = df_simu,
  chains = 4,
  warmup = 1000,
  iter = 5000,
  thin = 1,
  cores = 4,
  prior = prior_02,
  sample_prior = "only"
)
pp_check(model_ppc_02, type = "stat", stat = "min", prefix = "ppd")
pp_check(model_ppc_02, type = "stat", stat = "max", prefix = "ppd")
pp_check(model_ppc_02, type = "stat", stat = "mean", prefix = "ppd")
```

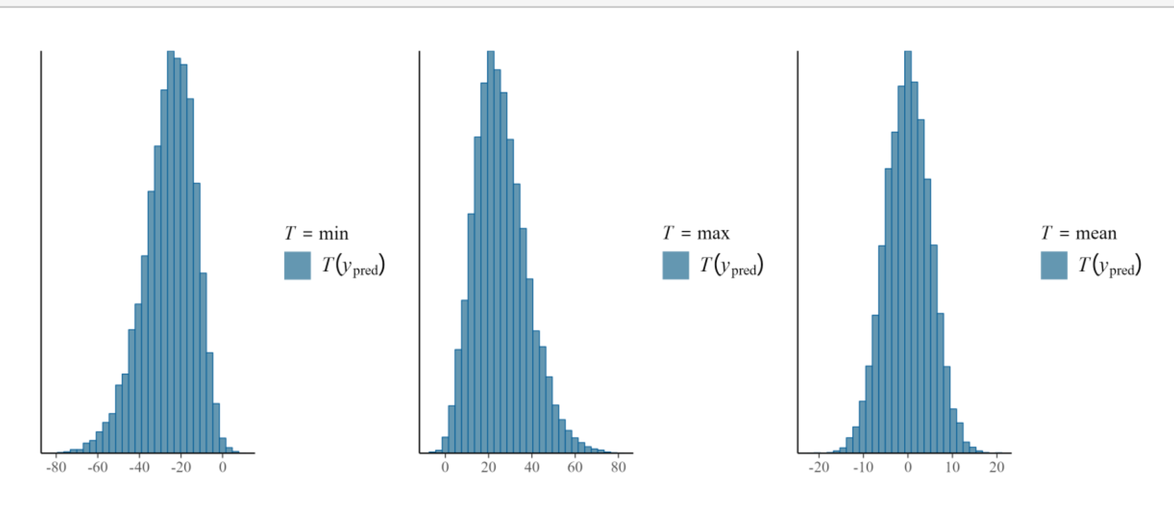



图 4 基于先验 prior_02 生成的每一套预测数据（y(pred)）的最小值、最大值和均值的分布

经过同样的步骤，我们可以获得基于先验 prior_02 生成的 16,000 套模拟观测数据（y(pred)）的最小值、均值和最大值的分布（图 4）。结果显示模拟观测数据的最小值大部分落在[-60, 0]，均值大部分落在了[-10, 10]，最大值大部分落在[0, 60]。与 prior_01 的先验预测结果（图 3）相比，prior_02 预测的结果更多落在了[-50, 50]内，能够相对更合理的描述我们的先验知识。最后需要注意的是，确定合理的先验分布通常需要重复多次类似的过程，且并不是先验分布的信息越强就越好，它需要合理反映研究者的先验知识。

1.6 模型拟合

在确定先验后，我们可以开始拟合模型，该模型包括所有可能的随机截距、随机斜率和它们之间的相关，并通过 *prior=prior_02* 为模型设定先前确定的先验。此外，我们设定 *sample_prior ="yes"*，以

此保留仅依靠先验信息获得的采样（即先验分布的采样）和同时依靠先验和数据似然值信息获得的采样（即后验分布的采样），这两种采样将会被用于计算贝叶斯因子（详见 3.5）。具体代码如下：

```
model_4 <- brm(
  LPP ~ group * type + (1 + type | subj),
  data = df_simu,
  chains = 4,
  warmup = 1000,
  iter = 5000,
  thin = 1,
  cores = 4,
  prior = prior_02,
  sample_prior = "yes"
)
summary(model_4)
# Group-Level Effects:
#   ~subj (Number of levels: 80)
#                         Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
# sd(Intercept)               2.16      0.18     1.85     2.54 1.00     1112     2707
# sd(type2M1)                 1.81      0.16     1.53     2.15 1.00     2626     5004
# cor(Intercept,type2M1)      0.31      0.11     0.10     0.51 1.00     2082     3882
#
# Population-Level Effects:
#                         Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
# Intercept                   0.67      0.24     0.18     1.13 1.01      656     1246
# group2M1                    0.27      0.49    -0.70     1.21 1.00      610     1141
# type2M1                     6.48      0.21     6.07     6.90 1.00     1569     3145
# group2M1:type2M1            0.00      0.43    -0.83     0.87 1.00     1247     2556
```

完成模型拟合后，我们可以借助后验预测检验（posterior predictive checking）检查模型对实际观测数据的“描述充分性”（“descriptive adequacy”，Gelman et al.，2013）。与先验预测检验（3.3 小节）相似，后验预测检验也可以根据参数值生成对应的模拟数据，区别在于，后验预测检验模拟过程使用的参数来自参数后验分布。我们同样可以使用 *brms* 的 *pp_check*（）函数来绘制后验预测检验的结果：

```
## 后验预测检验
pp_check(model_4, type = "stat_grouped", group = "type")
```

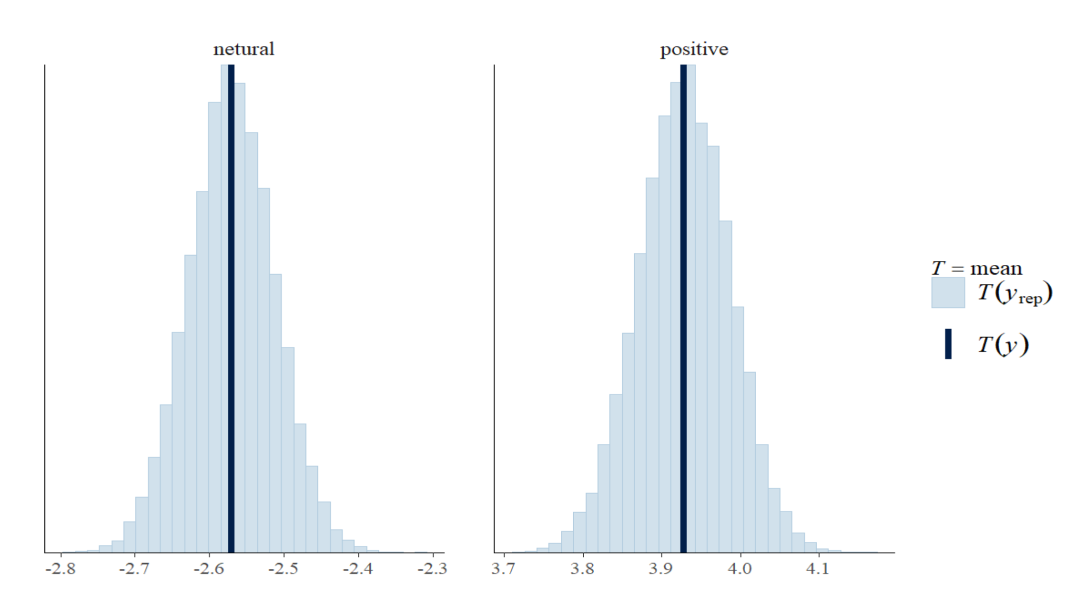


图 5 后验预测分布图（黑色线条 T（y）为观测数据的均值；浅蓝色部分 T（yrep）为后验预测数据的均值的分布）

图 5 展示了观测数据的均值（黑色线条为中性图片和积极图片类型 LPP 的均值）和依据模型生成的预测数据（即后验预测分布）的均值的分布（浅蓝色为模型预测的数据均值）。如图所示，大部分后验预测数据的均值与观测数据的均值接近，说明我们构建的模型较好地描述了观测数据。如果观测数据的统计量和模拟生成的预测数据的统计量间存在较大的差异，研究者则需要慎重考虑构建的模型是否忽略了某些重要因素，如未使用合适的残差分布或链接函数等（Schad et al.，2021）。

1.7 统计推断：贝叶斯因子

在实际应用中，研究者通常感兴趣的是实验条件之间是否存在差异，例如本研究中的组别（组间因素）和图片类型（被试内因素）是否会对 LPP 产生影响。在贝叶斯分析框架下，我们可以基于后验分布进行贝叶斯推断（Bayesian inference）和假设检验。进行假设检验的常见的方式之一是计算贝叶斯因子（Bayes factor）。贝叶斯因子可以量化基于不同的假设或模型（例如零假设和备择假设）观测到当前数据的似然比值（Rouder et al.，2018），研究者可以根据其数值大小推断观测到的数据更可能支持哪种假设或模型。一种流行的计算贝叶斯因子的方法是 Savage-Dickey density ratio，该方法通过计算在参数取值为 0 时两个模型的似然值的比值来估算贝叶斯因子，详见 Wagenmakers 等（2010）。通过 brms 提供的 *hypothesis()* 函数可以实现该贝叶斯因子的计算。

首先，利用 *model_4* 检验图片类型是否会对 LPP 产生影响：

```
# 组内比较
result <- hypothesis(model_4, "type2M1=0")
result$hypothesis$Evid.Ratio # 查看贝叶斯因子 BF_01
[1] 1.629488e-15
```

其中，*"type2M1=0"* 表示图片类型的零假设，即 LPP 在积极图片与中性图片上的差异为 0。*hypothesis$Evid.Ratio* 描述了贝叶斯因子 BF_{01}，表示基于零假设（H_0）观测到当前数据的似然与基于备择假设（H_1）观测到当前数据的似然的比值。为了依据 BF_{01} 进行统计推断，研究者通常需要预先设定贝叶斯因子的判断标准（胡传鹏等，2018）。根据 Lee and Wagenmakers' Scale 的贝叶斯因子判定标准 (Lee & Wagenmakers, 2014)，可以将BF_{01}支持零假设的证据强调分为十类，如表 2。上述结果显示，$BF_{01} = 1.63 * 10^{-15}$，远远小于 1/100，说明存在极强的证据支持图片类型的效应存在的备择假设，因此我们推断 LPP 的波幅在不同图片类型之间存在差异。

表 2 贝叶斯因子决策标准

贝叶斯因子，BF_{01}	**解释标准**
> 100	极强的证据支持 H_0
30 ~ 100	非常强的证据支持 H_0
10 ~ 30	较强的证据支持 H_0
3 ~ 10	中等程度的证据支持 H_0
1 ~ 3	较弱的证据支持 H_0
1	没有证据
1/3 ~ 1	较弱的证据支持 H_1
1/10 ~ 1/3	中等程度的证据支持 H_1
1/30 ~ 1/10	较强的证据支持 H_1
1/100 ~ 1/30	非常强的证据支持 H_1
< 1/100	极强的证据支持 H_1

注：该表改编自胡传鹏等，（2018），源引用于(Lee & Wagenmakers, 2014)。

接下来，我们检验 LPP 在不同组别间的差异是否为 0：

```
# 组间比较
result <- hypothesis(model_4, "group2M1=0")
result$hypothesis$Evid.Ratio # 查看贝叶斯因子 BF_01
[1] 8.865514
```

"group2M1=0" 表示组别（group）的零假设，即抑郁组与对照组被试在 LPP 上的表现差异为 0。结果显示，BF_{01}= 8.90 > 3，这表明存在中等强度的证据支持组别的零假设，从而我们推断组别之间的 LPP 波幅不存在差异。

此外，我们还可以检验交互作用是否为 0:

```
result <- hypothesis(model_4, "group2M1:type2M1=0")
result$hypothesis$Evid.Ratio # 查看贝叶斯因子 BF_01
[1] 11.74594
```

"group2M1:type2M1=0" 表示组别（group）和图片类型（type）交互作用的零假设，即抑郁组中积极图片和中性图片 LPP 波幅差异与对照组中积极图片和中性图片 LPP 波幅差异的差异（即差的差）为 0。结果显示，$BF_{01} = 11.75 > 10$，这表明存在较强的证据支持不存在组别和图片类型的交互作用。

最后，我们可以根据 *model_4* 参数分析的结果（固定效应）和贝叶斯因子分析的结果报告：“贝叶斯混合效应模型的结果显示，存在极强的证据表明，积极图片诱发的 LPP 波幅高于中性图片，$\beta_{图片类型} = 6.48$，$BF_{01} = 1.63 * 10^{-15}$；同时，存在较强的证据表明，两组的 LPP 波幅不存在显著差异，$\beta_{组别} = 0.27$，$BF_{01} = 8.90$；且存在中等的证据表明图片类型和组别不存在交互作用，$\beta_{图片类型*组别} = 0$，$BF_{01} = 11.75$”。

4 小结

本文首先介绍了贝叶斯混合效应模型在处理数据层级结构和提供更直观的统计结果等方面的独特优势，并概述了其基本原理。更重要的是，本文结合模拟数据和代码示例，介绍了固定效应和随机效应， 如何构建贝叶斯混合效应模型，如何利用先验预测检验设置合理的先验分布，以及如何借助贝叶斯因子进行统计推断。

本文主要关注了贝叶斯线性混合效应模型，它是其他扩展模型的基础。一类常用的扩展模型是贝叶斯广义线性混合效应模型（Bayesian Generalized Linear Mixed-effects Models），它能够更好地处理心理学实验中其他常见的因变量，比如二元选择、反应时、阅读时间、李克特评分等（Dixon，2008；Rousselet & Wilcox，2020）。具体来说，研究者可以在贝叶斯广义线性混合效应模型中为因变量设定更合理的残差分布（residual distribution）和链接函数（link function），进而获得更准确的估计结果（Vuorre，2017）。例如，对于二项选择（如正确标记为 1，错误为 0）的因变量，*brm()*函数可以通过参数 *family= bernoulli("logit")*设置模型的残差分布为伯努利（Bernoulli）分布且使用 logit 链接函数（Bürkner，2017）。再比如对于反应时，研究者可以通过 *family=“lognormal”* 为其设置 log 转换（Wagenmakers & Brown，2007）。此外，贝叶斯分层模型还可以应用于复杂的认知模型，比如信号检测论、漂移扩散模型、强化学习模型、多项式加工树模型等（Ahn et al.，2017；Schad et al.，2020；Tso et al.，2021）。这些分析可以通过 R 语言的工具包（如 hBayesDM）、Stan 和 JAGS 等其他工具来完成（Ahn et al.，2017；Kruschke，2014）。总的来说，贝叶斯混合效应模型凭借其强大的灵活性，能够很好地适用于多样的心理学实验研究中。

贝叶斯混合效应模型的诸多优点使其逐渐流行于心理学研究的数据分析之中（Bono et al.，2021）。但不可否认的是，研究者在实践中使用这种分析方法时仍会面临一些困难。比如，在初学贝叶斯混合效应模型时，研究者可能并不清楚如何设置合理的先验，进而会采用工具包提供的默认先验。需要注意的是，这些默认先验并不一定适用于所有的研究，而使用不合适的先验可能造成模型难以收敛等潜

在问题（Veenman et al.，2022）。如 3.5 小节所介绍，研究者可以根据以往的研究和自己的知识自定义先验，并且可以借助先验预测检验来审查这些先验的合理性。其次，当使用贝叶斯因子进行假设检验时，研究者一般无法通过单个模型获得所有感兴趣的效应。例如 *model_4* 结果只包含了图片类型和组别的主效应，以及他们的交互作用。相应地，我们也只能计算这三个效应的贝叶斯因子。如果研究者同时对其他效应（如简单效应）感兴趣，则需要构建新的模型并设置不同的对比编码（contrast coding）来检验不同的假设（Schad et al.，2020）。最后，拟合贝叶斯混合效应模型相对更加耗时。随着数据量的变大、模型复杂程度的提升、以及采样迭代数量的增长，贝叶斯混合效应模型的拟合时间也会相应地呈指数性增长。对此，研究者可以通过设置合理的先验和使用高性能的计算机提升模型的拟合速度。

本文借助心理学研究示例，通过具体代码向读者展示了如何使用贝叶斯混合效应模型进行数据分析。尽管我们尝试涵盖尽可能多的内容，但我们介绍的也只是贝叶斯混合效应模型的冰山一角。随着开放科学（Open Science）的兴起（胡传鹏等，2016；Nosek，2019；Open Science Collaboration，2015；Jin et al., 2023），越来越多的文章公开其数据和代码，这无疑为研究者提供了更多结合具体实例的贝叶斯混合效应模型的应用教程。此外，我们也鼓励希望对该方法有更加深入和全面了解的读者阅读贝叶斯分析相关的优秀书籍（Gelman et al.，2013；Kruschke，2014；Lambert，2018；McElreath，2020）。希望研究者可以对叶斯混合效应模型有更加深入的了解，能够广泛地将其应用于多样的心理学研究之中。

参考文献

胡传鹏, 孔祥祯, Eric-Jan Wagenmakers, Alexander Ly, 彭凯平. (2018). 贝叶斯因子及其在 JASP 中的实现. *心理科学进展*, *26*(06), 951–965.

胡传鹏, 王非, 过继成思, 宋梦迪, 隋洁, 彭凯平. (2016). 心理学研究中的可重复性问题:从危机到契机. *心理科学进展*, *24*(09), 1504–1518.

Ahn, W.-Y., Haines, N., & Zhang, L. (2017). Revealing Neurocomputational Mechanisms of Reinforcement Learning and Decision-Making With the hBayesDM Package. *Computational Psychiatry*, *1*(0), 24–24.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278.

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). *Parsimonious mixed models.* ArXiv.

Bono, R., Alarcón, R., & Blanca, M. J. (2021). Report quality of generalized linear mixed models in psychology: A systematic review. *Frontiers in Psychology*, *12*, 666182.

Bürkner, P.-C. (2017). Brms: A package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1-28.

Bürkner, P.-C. (2018). Advanced Bayesian multilevel modeling with the R package brms. *The R Journal*, *10*(1), 395–411.

Cohen, J. (1994). The earth is round (p < .05). *American Psychologist*, *49*(12), 997–1003.

DeBruine, L. (2021). *faux: Simulation for Factorial Designs R Package version 1.1.0*. Zenodo.

Dixon, P. (2008). Models of accuracy in repeated-measures designs. *Journal of Memory and Language*, *59*(4), 447–456.

Eager, C., & Roy, J. (2017). *Mixed Effects Models are Sometimes Terrible*. ArXiv.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian Data Analysis*. Chapman and Hall/CRC.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 457-472.

Gelman, A., Vehtari, A., Simpson, D., Margossian, C. C., Carpenter, B., Yao, Y., Kennedy, L., Gabry, J., Bürkner, P.-C., & Modrák, M. (2020). Bayesian Workflow. 2020.

Haaf, J. M., & Rouder, J. N. (2017). Developing constraint in Bayesian mixed models. *Psychological Methods*, *22*(4), 779–798.

Jin, H., Wang, Q., Yang, Y.-F., Zhang, H., Gao, M. (Miranda), Jin, S., Chen, Y. (Sharon), Xu, T., Zheng, Y.-R., Chen, J., Xiao, Q., Yang, J., Wang, X., Geng, H., Ge, J., Wang, W.-W., Chen, X., Zhang, L., Zuo, X.-N., & Chuang-Peng, H. (2023). The Chinese Open Science Network (COSN): Building an open science community from scratch. *Advances in Methods and Practices in Psychological Science*, *6*(1).

Judd, C. M., Westfall, J., & Kenny, D. A. (2017). Experiments with more than one random factor: designs, analytic models, and statistical power. *Annual Review of Psychology*, *68*(1), 601–625.

Kruschke, J. K. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan*. Academic Press.

Kruschke, J. K., & Liddell, T. M. (2018). Bayesian data analysis for newcomers. *Psychonomic Bulletin & Review*, *25*(1), 155–177.

Lambert, B. (2018). *A student's guide to Bayesian statistics*. SAGE.

Lewandowski, D., Kurowicka, D., & Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis*, *100*(9), 1989–2001.

Magezi, D. A. (2015). Linear mixed-effects models for within-participant psychology experiments: An introductory tutorial and free, graphical user interface (LMMgui). *Frontiers in Psychology*, *6*(JAN), 1–7.

Makowski, D., Ben-Shachar, M., & Lüdecke, D. (2019). bayestestR: Describing effects and their uncertainty, existence and significance within the Bayesian framework . *Journal of Open Source Software*, *4*(40), 1541.

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language*, *94*, 305–315.

McElreath, R. (2020). *Statistical rethinking: a Bayesian course with examples in R and Stan*. CRC Press.

Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., & Wagenmakers, E.-J. (2016). The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin & Review*, *23*(1), 103–123.

Morey, R. D., Rouder, J. N., Pratte, M. S., & Speckman, P. L. (2011). Using MCMC chain outputs to efficiently estimate Bayes factors. *Journal of Mathematical Psychology*, *55*(5), 368–378.

Nosek, B. A. (2019). The rise of open science in psychology, a preliminary report. *Center for Open Science*.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716.

Plummer, M. (2023). Simulation-Based Bayesian Analysis. *Annual Review of Statistics and Its Application*, *10*.

Rouder, J. N., Haaf, J. M., & Vandekerckhove, J. (2018). Bayesian inference for psychology, part IV: parameter estimation and Bayes factors. *Psychonomic Bulletin & Review*, *25*(1), 102–113.

Rousselet, G. A., & Wilcox, R. R. (2020). *Reaction Times and other Skewed Distributions: Problems with the Mean and the Median*. BioRxiv.

Roy, V. (2020). Convergence diagnostics for Markov Chain Monte Carlo. *Annual Review of Statistics and Its Application*, *7*(1), 387–412.

Schad, D. J., Betancourt, M., & Vasishth, S. (2021). Toward a principled Bayesian workflow in cognitive science. *Psychological Methods*, *26*(1), 103–126.

Schad, D. J., Vasishth, S., Hohenstein, S., & Kliegl, R. (2020). How to capitalize on a priori contrasts in linear (mixed) models: A tutorial. *Journal of Memory and Language*, *110*, 104038.

Seaman, J. W., Seaman, J. W., & Stamey, J. D. (2012). Hidden dangers of specifying noninformative priors. *The American Statistician*, *66*(2), 77–84.

Singmann, H., & Kellen, D. (2019). An Introduction to Mixed Models for Experimental Psychology. In *New Methods in Cognitive Psychology* (pp. 4–31). Routledge.

Sorensen, T., & Vasishth, S. (2016). Bayesian linear mixed models using Stan: A tutorial for psychologists, linguists, and cognitive scientists. *The Quantitative Methods for Psychology*, *12*(3), 175–200.

Tso, I. F., Taylor, S. F., & Johnson, T. D. (2021). Applying hierarchical bayesian modeling to experimental psychopathology data: An introduction and tutorial. *Journal of Abnormal Psychology*, *130*(8), 923–936.

Tuerlinckx, F., Rijmen, F., Verbeke, G., & De Boeck, P. (2006). Statistical inference in generalized linear mixed models: A review. *British Journal of Mathematical and Statistical Psychology*, *59*(2), 225–255.

Veenman, M., Stefan, A., & Haaf, J. M. (2022). *Bayesian Hierarchical Modeling: An Introduction and Reassessment*. PsyArXiv.

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P.-C. (2021). Rank-normalization, folding, and localization: An improved $\hat{R}$ for assessing convergence of MCMC (with Discussion). *Bayesian Analysis*, *16*(2), 667-718.

Venables, W. N., & Ripley, B. D. (2002). *Modern Applied Statistics with S*. New York: Springer.

Vuorre, M. (2017). *Bayesian Estimation of Signal Detection Models*.

Wagenmakers, E.-J., & Brown, S. (2007). On the linear relation between the mean and the standard deviation of a response time distribution. *Psychological Review*, *114*(3), 830–841.

Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the Savage–Dickey method. *Cognitive Psychology*, *60*(3), 158–189.

Bayesian Mixed-effects Models: A Primer with brms

Wanke Pan[1] Xiujuan Wen[2] Haiyang Jin[3]

([1] School of Psychology, Nanjing Normal University, Nanjing 210097, China; [2] Department of Neurology, The Affiliated Brain Hospital of Guangzhou Medical University, Guangzhou 510370, China; [3]Department of Psychology, Division of Science, New York University Abu Dhabi, PO Box 129188, Saadiyat Island, Abu Dhabi, United Arab Emirates)

Corresponding author: Haiyang Jin, E-mail: haiyang.jin@outlook.com

Abstract: Compared to the traditional statistical methods, Bayesian linear mixed-effects modeling (BLMM) has a great number of advantages in dealing with the hierarchical structures underlying datasets and providing more intuitive statistical results. These advantages together popularize BLMM in psychological and other field research. However, there is still a lack of tutorials on the practical applications of BLMM in psychology studies in Chinese. Therefore, we first briefly introduced the basic concepts and rationales of BLMM. Then we employed a simulated dataset to demonstrate how to understand fixed effects and random effects, and how to use the popular brms R package to specify models for BLMM based on the experimental design. We additionally covered the procedure of pre-specifying priors with prior predictive checks, and the steps of performing hypothesis testing using the Bayes Factor. BLMM, with its extensions such as Generalized BLMM, has great flexibility and capability, they can and should be applied in various psychology research.

Key words: Bayesian, linear mixed-effects modeling, hierarchical models, Bayes Factor, brms